Assignment: Author Prediction

Problem Statement -Given the content, your task is to predict the author.

About Dataset

> File 1 - content_author_assignment_train.csv
>
> File 2 - content_author_assignment_test.csv

The train file for any analysis and training The test file can solely be used for prediction.

Columns - content, author

Evaluation criteria -cross entropy loss

Loading required libraries

> Importing support utils file for preprocessing and feature extraction as well

In [1]:
```
import csv
import sys
import pandas as pd
from sklearn import svm
from xgboost import XGBClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import SGDClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.multiclass import OneVsRestClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
from sklearn.model_selection import GridSearchCV,cross_val_score,train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_curve, auc, log_loss

sys.path.append("../")
# other local imports
from utils.nlp_utils import *
from utils.text_analysis import FeatureEng
```

Initializing Global Parameters to be used during Classifiction

In [2]:
```
'''Global Parameters'''
features = FeatureEng
LE = LabelEncoder()
tf_idf_vectorizer = TfidfVectorizer(lowercase=True)
MNB_classifier = MultinomialNB(alpha=0.5)
xgBoost = XGBClassifier()
rest_classifier = OneVsRestClassifier(SGDClassifier(loss='log', alpha=0.001,penalty='l2'), n_jobs=-1)
rf = RandomForestClassifier(random_state=3)

model_dict = {'XGBoost Classifier' : xgBoost,
              'Multinomial Naive Bayes' : MNB_classifier,
              'OneVsRest Classifier': rest_classifier,
              'Random Forest': rf,
              'AdaBoost': AdaBoostClassifier(random_state=3),
              'K Nearest Neighbor': KNeighborsClassifier(),
              'Stochastic Gradient Descent' : SGDClassifier(random_state=3, loss='log')}
```

Reading Training Data into Dataframe

In [3]:
```
df = pd.read_csv("../data/content_author_assignment_train.csv")
df.head()
```

Out[3]:

	content	author
0	Fourth time this season, heavy rains paralysed...	The Quint
1	The BJP has made a significant gain of 11 addi...	PTI
2	Rajya Sabha saw two disruptions on Thursday al...	FP Staff
3	Senior IPS officer Subodh Jaiswal, who had bee...	The Quint
4	The government, on 27 February, announced a si...	PTI

Preprocessing content column by applying basic cleaning pipeline

In [4]:
```
df['cleaned_content'] = df.content.apply(lambda x: features.clean_text(x))
df['cleaned_content'] = df.content.apply(lambda x: features.remove_stopwords(x))
# mapped author_name to numbers
df['author_id'] = LE.fit_transform(df['author'])
print("\n Shape of dataframe passed:" ,df.shape)
df.head()
```

```
 Shape of dataframe passed: (712, 4)
```

Out[4]:

	content	author	cleaned_content	author_id
0	Fourth time this season, heavy rains paralysed...	The Quint	Fourth time season, heavy rains paralysed city...	4
1	The BJP has made a significant gain of 11 addi...	PTI	The BJP made significant gain 11 additional se...	1
2	Rajya Sabha saw two disruptions on Thursday al...	FP Staff	Rajya Sabha saw two disruptions Thursday alrea...	0
3	Senior IPS officer Subodh Jaiswal, who had bee...	The Quint	Senior IPS officer Subodh Jaiswal, working cou...	4
4	The government, on 27 February, announced a si...	PTI	The government, 27 February, announced signifi...	1

In [5]:
```
def get_author_map(df):
    author_map = {}
    list_of_author = list(df.author.unique())
    for i in list_of_author:
        value =df.loc[df['author'] == i, 'author_id'].iloc[0]
        author_map[i] = value
    return author_map

author_map = get_author_map(df)
author_map
```

Out[5]:
```
{'The Quint': 4,
 'PTI': 1,
 'FP Staff': 0,
 'Press Trust of India': 2,
 'Scroll Staff': 3}
```

Split data into train and test

> For evaluation purposes

In [6]:
```
X_train, X_test, y_train, y_test = train_test_split(df['cleaned_content'],
                                                    df['author_id'], test_size=0.3, random_state=42)
print("X_train shape: ",X_train.shape)
print("X_test shape: ",X_test.shape)
print("y_train shape: ",y_train.shape)
print("y_test shape: ",y_test.shape)

X_train shape:  (498,)
X_test shape:  (214,)
y_train shape:  (498,)
y_test shape:  (214,)
```

Converting Text into features

> For this we will be using TF-IDF vectorizer.

In [7]:
```
X_train_vectorized = tf_idf_vectorizer.fit_transform(X_train)
X_test_vectorized = tf_idf_vectorizer.transform(X_test)
```

Training the learning algorithm

In [8]:
```
def model_score_df(model_dict):
    model_name, ac_score_list, p_score_list, r_score_list, f1_score_list,log_loss_list = [],[], [], [], [], []
    for k,v in model_dict.items():
        model_name.append(k)
        v.fit(X_train_vectorized,y_train)
        predictions = v.predict(X_test_vectorized)
        pred = v.predict_proba(X_test_vectorized)
        ac_score_list.append(accuracy_score(y_test, predictions))
        p_score_list.append(precision_score(y_test, predictions, average='macro'))
        r_score_list.append(recall_score(y_test, predictions, average='macro'))
        f1_score_list.append(f1_score(y_test, predictions, average='macro'))
        log_loss_list.append(log_loss(y_test, pred))
    model_comparison_df = pd.DataFrame([model_name, ac_score_list, p_score_list, r_score_list, f1_score_list, log_loss_list]).T
    model_comparison_df.columns = ['model_name', 'accuracy_score', 'precision_score', 'recall_score', 'f1_score', 'log_loss']
    model_comparison_df = model_comparison_df.sort_values(by='log_loss', ascending=False)

    return model_comparison_df
```

Evaluation

In [9]:
```
model_score_df(model_dict)
```

```
/home/shivani/.local/lib/python3.8/site-packages/xgboost/sklearn.py:888: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)

[13:02:24] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

/home/shivani/.local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/home/shivani/.local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Out[9]:

	model_name	accuracy_score	precision_score	recall_score	f1_score	log_loss
5	K Nearest Neighbor	0.38785	0.431872	0.342672	0.324424	6.490018
1	Multinomial Naive Bayes	0.46729	0.240255	0.296413	0.240214	1.53241
2	OneVsRest Classifier	0.546729	0.41929	0.402452	0.382315	1.22977
0	XGBoost Classifier	0.626168	0.58422	0.590688	0.58433	1.1889
4	AdaBoost	0.406542	0.593643	0.34181	0.331047	1.13784
6	Stochastic Gradient Descent	0.551402	0.537562	0.467176	0.472397	1.08395
3	Random Forest	0.546729	0.501457	0.393524	0.362865	1.0296

> **Observation:** XGBoost classifier is giving overall better results than the rest of the classifiers. So for prediction, we will proceed with xgboost classifier.

Creating inference function

In [10]:
```
unseen_data = pd.read_csv('../data/content_author_assignment_test.csv')
unseen_data_raw = unseen_data.copy()
unseen_data.head()
```

Out[10]:

	content	author
0	The Shiv Sena will abstain from voting on the...	Press Trust of India
1	Disgruntled BJP leader Shatrughan Sinha, who h...	PTI
2	The Congress would emerge as the "number one ...	PTI
3	Former Nationalist Congress Party leader Tariq...	Scroll Staff
4	Janata Dal (United) president Nitish Kumar on ...	The Quint

Preparing unseen data for inferencing

In [11]:
```
unseen_data['cleaned_content'] = unseen_data.content.apply(lambda x: features.clean_text(x))
unseen_data['cleaned_content'] = unseen_data.content.apply(lambda x: features.remove_stopwords(x))
print(unseen_data.info())
print(f' Shape of dataframe passed:' ,unseen_data.shape)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 855 entries, 0 to 854
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   content          855 non-null    object
 1   author           855 non-null    object
 2   cleaned_content  855 non-null    object
dtypes: object(3)
memory usage: 20.2+ KB
None
 Shape of dataframe passed: (855, 3)
```

Inferencing function

In [14]:
```
def make_inference(content):
    query_vector = tf_idf_vectorizer.transform([content])
    predicted_author = xgBoost.predict(query_vector)
    author= list(author_map.keys())[list(author_map.values()).index(predicted_author)]
    return author
```

In [15]:
```
author_pred = make_inference(unseen_data['cleaned_content'][2])
print("Actual value: ",unseen_data_raw['author'][2])
print("Predicted Value:", author_pred)

Actual value:  PTI
Predicted Value: PTI
```

In [16]:
```
predicted_author = []
for content in unseen_data['cleaned_content']:
    pred = make_inference(content)
    predicted_author.append(pred)
```

In [17]:
```
author_predictions = pd.DataFrame({'Content text':unseen_data['content'],'Author':unseen_data['author'],'Predicted Author':predicted_author})
```

In [18]:
```
author_predictions.head(10)
```

Out[18]:

	Content text	Author	Predicted Author
0	The Shiv Sena will abstain from voting on the...	Press Trust of India	PTI
1	Disgruntled BJP leader Shatrughan Sinha, who h...	PTI	The Quint
2	The Congress would emerge as the "number one ...	PTI	PTI
3	Former Nationalist Congress Party leader Tariq...	Scroll Staff	Scroll Staff
4	Janata Dal (United) president Nitish Kumar on ...	The Quint	The Quint
5	The Madras High Court on Tuesday directed that...	The Quint	The Quint
6	The Aam Aadmi Party announced on Tuesday its ...	PTI	PTI
7	Suspended Congress leader Mani Shankar Aiyar s...	The Quint	The Quint
8	After his "Internet in the Mahabharata era" re...	PTI	The Quint
9	"Marriage does not mean that the woman is all ...	The Quint	The Quint

In [ ]:
```
# saving prediction dataframe to csv
author_predictions.to_csv('../data/author_predictions.csv',index=False)
```

In [ ]:
```

```